TP2 :
Word Embeddings (Plongement lexicaux)

Chaimae EL HOUJJAJI

31 janvier 2022

1 Instructions et Context

1.1 Presentation des corpus

Nous avons 2 corpus sur deux thèmes différents : un sur le domaine médical et un sur le domaine de la presse. Les deux fichiers sont :

- `QUAERO_FrenchMed_traindev.ospl`. Il s'agit d'un corpus en français dans le domaine médical qui est composé de **3091 lignes et 51896 mots** (donnée récupéré en utilisant la commande `wc`)
- `QUAERO_FrenchPress_traindev.ospl`. Il s'agit d'un corpus en français dans le domaine de la press qui est composé de **38548 lignes et 1252121 mots**

Ainsi le corpus sur la presse contient environ **24 fois plus de mots** que le corpus médical et environ 12 fois plus de lignes. Ainsi le corpus médical est un corpus de petite taille par rapport au corpus non médical.

1.2 Embeddings de mots

En 2013, une nouvelle méthode de vectorisation de texte appelée words embeddings est apparu dans le domaine du NLP.
Ces nouvelles techniques de vectorisation de texte se distinguent des précédentes méthodes de vectorisation de mots. En effet, elles ont réussi à conserver la similarité sémantique entre les mots, ce qui signifie que ces vecteurs peuvent reconnaître le sens d'un mot et déterminer sa similarité avec d'autres. C'est ce que nous allons testé dans ce projet et nous allons également comparer ces approches.

1.3 Packages utilisés

Afin de faire ces plongements lexicaux (word embeddings), différentes méthodes sont apparues progressivement :

- 2013 : `Word2VC` par Thomas Mikolov chez Google.
- 2014 : `GloVe` par Jeffrey Pennington à Stanford.
- 2016 : `fastText` par Piotr Bojanowski chez Facebook.

Dans le cadre de notre projet, nous étudierons les approches `Word2VC` et `fastText`. Ces deux approches propose deux scénarios différents pour la formation du réseau de neurones :

- **CBOW** ou "continuous bag-of-words" prédit le mot en fonction d'un certain contexte . Cette méthode fonctionne bien pour les mots courants, mais moins bien pour les mots rares.
- **Skip-Gram** prédit le contexte en fonction du mot. Elle fonctionne bien avec une petite quantité de données d'entraînement et représente correctement les mots ou les phrases rares.

Nous allons donc utiliser les approches `word2vec` (Cbow, skipgram) et `fasttext` (Cbow).

1.4 Objectif du projet

L'objectif de ce projet est de pouvoir comparer différentes approches de "word embeddings" appliqués sur des corpus de taille différentes et sur des domaines différents. Pour cela, nous allons dans un premier temps aprendre les embeddings de mots pour chaque approche. Cela nous donnera au final 6 embeddings.
Puis nous allons étudier la similarité sémantique, c'est à dire trouver les 10 mots les plus proches d'un mot donné en s'appuyant sur le calcul de similarité cosinus. Ces mots sont : `patient`, `traitement`, `maladie` et `jaune`. On obtient ainsi pour chacun de

ces mots, leurs 10 mots les plus proches ainsi que leur score de similarité (entre 0 et 1).

Pour ce projet, il aurait été utile d'effectuer une partie de préparation et de nettoyage des données mais nous préférons nous concentrer sur la comparaison des différentes approches. Toutefois nous décidons quand même lors de l'étape de chargement du corpus, d'enlever toutes les ponctuations ainsi que d'enlever les mots de taille égale à 1 suite à l'obtention une première fois de termes de ponctuation lors de la recherche des 10 mots les plus proches.
Pour faciliter la lecture de ce rapport, j'ai créé un github pour y déposer tous les documents nécessaires et utilisés dans ce projet. Par la suite, des liens seront insérés à chaque partie pour faire référence aux scripts python utilisés et présent dans ce github.

1.5 Paramètres utilisés

Un certain nombre de paramètres ont été utilisés pour entrainer les modèles :
- `min_count = 1`, je considère tout le corpus (autrement dit, je ne considère pas les mots qui reviennent moins de 1 fois, mais cela revient à considérer tout le corpus).
- `dim = 100`, c'est à dire que les mots sont codés dans des vecteurs de taille 100.
- `epochs=50`, on regarde chaque mot du corpus 50 fois durant l'étape du "training".

2 Résultats

2.1 Aprentissage des embeddings

Nous avons donc au total appris 6 embeddings :
- Word2vec avec CBOW sur le corpus médical
- Word2vec avec CBOW sur le corpus non-médical (presse)
- Word2vec avec Skip-Gram sur le corpus médical
- Word2vec avec Skip-Gram sur le corpus non-médical (presse)
- FastText avec CBOW sur le corpus médical
- FastText avec CBOW sur le corpus non-médical (presse)

Le script d'entrainement des modèles : Entrainement.py est disponible sur le github. Pour visualiser l'ensemble des mots les plus proches pour les 6 approches, un fichier texte a été généré afin de permettre de récapituler l'ensemble des mots et les scores de similarité : WordEmbeddings_Results.txt.

2.2 Comparaison des embeddings entrainés sur le même corpus

2.2.1 Corpus Médical

Nous allons dans un premier temps comparer les 10 mots les plus proches obtenues par comparaison des 3 approches (Word2vec CBOW, Word2vec SkipGram et FastText CBOW) sur le corpus médical afin de voir l'impact de l'approche sur un même corpus de petite taille et où les 4 mots cibles font partie du champ lexical du corpus. Les résultats sont résumés dans le tableau à la Figure 1.

patient						traitement						maladie						jaune					
w2v CBOW		w2v Sk		ft CBOW		w2v CBOW		w2v Sk		ft CBOW		w2v CBOW		w2v Sk		ft CBOW		w2v CBOW		w2v Sk		ft CBOW	
soigneusement	0,986	souffre	0,847	Patient	0,987	par	0,968	concomitant	0,790	Traitement	0,987	directement	0,986	thrombo	0,902	malade	0,980	LA	0,997	blanc	0,997	hexane	0,931
lorsqu	0,982	aptitude	0,847	patiente	0,961	exemple	0,968	début	0,788	Taaitement	0,978	ADN	0,985	grave	0,902	Maladie	0,972	DU	0,997	valeurs	0,996	goutte	0,923
allergique	0,979	modifier	0,842	parvient	0,948	positif	0,958	Arrêt	0,784	traitment	0,973	la	0,985	atteint	0,899	maldi	0,951	Etude	0,997	couleur	0,996	gêne	0,921
signe	0,979	devrait	0,829	appartient	0,945	arrêt	0,956	mois	0,783	Allaitement	0,973	rare	0,985	thromboembolique	0,891	malgré	0,919	EN	0,997	polyéthylène	0,996	titane	0,906
confirmé	0,977	conscient	0,828	maintient	0,945	diagnostic	0,956	expérimenté	0,782	allaitement	0,968	Crohn	0,982	Crohn	0,888	malaise	0,918	DE	0,997	méthyle	0,996	banane	0,904
devra	0,977	personnel	0,825	recevaient	0,928	confirmé	0,954	Traitement	0,781	étroitement	0,967	embolique	0,981	embolique	0,882	amantadie	0,885	DES	0,996	créatinine	0,996	Rhône	0,902
particulièrement	0,974	présente	0,824	patients	0,918	comité	0,953	antérieurs	0,775	évitement	0,965	groupe	0,981	considérée	0,882	maligne	0,877	syndrome	0,996	HDPE	0,996	perchlorure	0,899
Il	0,974	analgésique	0,822	gradient	0,915	antérieur	0,951	résultat	0,770	entrainement	0,951	prudence	0,980	malgré	0,872	malin	0,877	lamivudine	0,996	tumeur	0,996	Hygiène	0,898
agir	0,972	allergique	0,822	excipient	0,915	affections	0,950	définitif	0,769	département	0,949	fin	0,980	rare	0,870	intolérance	0,858	Lilly	0,996	phosphate	0,996	jeune	0,895
aptitude	0,971	remarquer	0,820	passant	0,915	TYSABRI	0,947	antérieur	0,769	recrutement	0,948	levure	0,980	moelle	0,853	Parkinson	0,852	32	0,995	points	0,996	Lane	0,895

FIGURE 1 – 10 Mots les plus proches de 4 mots : patient, traitement, maladie et jaune par 3 approches différentes pour le corpus médical

On remarque sur ce tableau plusieurs choses. Tout d'abord on remarque que l'approche fasttext CBOW réussi à trouver des mots qui ont un grand nombre de lettres en communs avec le mot cible, ce qui conduit à des résultats qui sont loins du mot cible et qui n'ont pas de rapport entre eux. On voit également l'importance de l'étape de lémmatization qui n'a pas été faite ici puisqu'en effet certains mots trouvés correspondent au même mot cible mais juste au pluriel, au féminin ou avec une majuscule en plus. (ex : patient - Patient - patiente patients). En ce qui concerne les approches Word2Vec CBOW et SkipGram, on remarque que parmi les 10 mots trouvés, il y en a en moyenne 3 ou 4 qui sont en lien avec le domaine médical. Ces approches semblent mieux fonctionner que FastText, toutefois il est difficile de bien mesurer l'avantage de l'une par rapport à l'autre bien que l'approche

Word2Vec SkipGram semple donner plus de mots en commun.

Nous avons donc essayé de visualiser cela graphiquement. Pour cela nous sommes donc passé d'une représentation des vecteurs avec 100 dimensions à une représentation en 2 dimensions grâce à une TSNE. (Une autre approche possible aurait été de faire une ACP). Le script pour réaliser cela est Viz_med_wordembeddings.py.

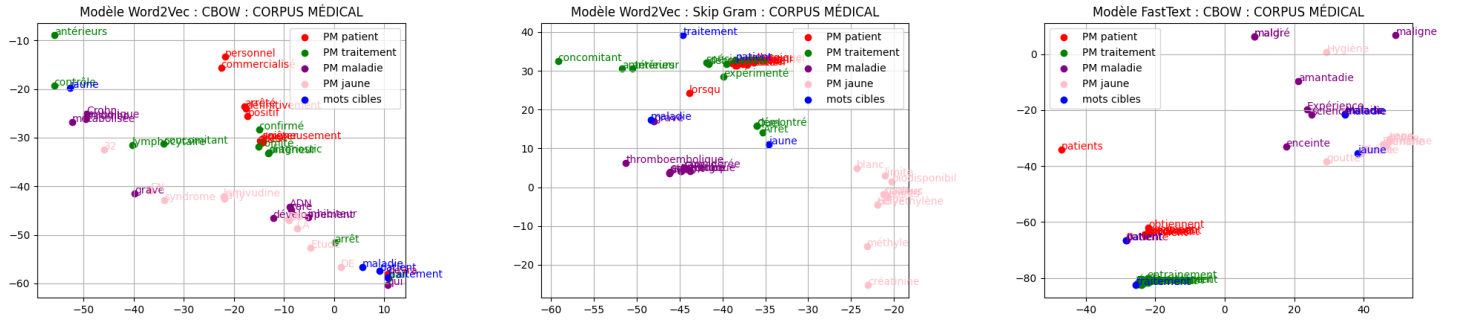

FIGURE 2 – Représentation graphique des mots les plus proches pour 3 approches sur un corpus médical.

D'après la Figure 2, on remarque que dans l'approche Word2Vec CBOW, le mot jaune se distingue bien des mots maladie, traitement et patient qui sont quant à eux regroupés. Toutefois on arrive pas à bien distinguer les clusters associés aux mots cibles et à leurs mots les plus proches.
Ces clusters sont toutefois bien plus visibles dans la représentation avec le modèle Word2Vec Skipgram et avec FastText.

2.2.2 Corpus Non-Médical (de la presse)

Nous allons ensuite comparer les 10 mots les plus proches obtenues par comparaison des 3 approches (Word2vec CBOW, Word2vec SkipGram et FastText CBOW) sur le corpus non médical qui est plus long mais où les mots cibles (patients, maladie, traitement, jaune) font partie d'un vocabulaire du domaine différent : le domaine médical.

patient						traitement						maladie						jaune					
w2v CBOW		w2v Sk		ft CBOW		w2v CBOW		w2v Sk		ft CBOW		w2v CBOW		w2v Sk		ft CBOW		w2v CBOW		w2v Sk		ft CBOW	
malaise	0,888	aise	0,956	patientent	0,945	poumon	0,933	alimentaire	0,897	retraitement	0,918	résolution	0,866	assurance	0,870	malade	0,787	Michael	0,972	maillot	0,948	brune	0,798
consensus	0,883	complément	0,953	impatient	0,940	coût	0,925	mépris	0,890	subitement	0,888	responsabilité	0,864	susceptible	0,782	maladies	0,767	Joey	0,965	Bastad	0,912	Jeune	0,783
souffle	0,881	clown	0,951	détient	0,926	contenu	0,922	financement	0,887	recrutement	0,884	puissance	0,860	bénéficie	0,761	malawite	0,700	Spadea	0,964	SaintMarin	0,897	Neptune	0,778
emprunt	0,877	insupportable	0,951	impatientent	0,909	lycée	0,918	destiné	0,884	bêtement	0,873	perspective	0,851	potentiel	0,761	malnutrie	0,693	maillot	0,962	Rafael	0,890	Saâdoune	0,756
diplôme	0,874	irréaliste	0,949	renient	0,903	cancer	0,917	potentiel	0,881	doctement	0,872	constitution	0,849	garantie	0,759	maladresse	0,672	Kahn	0,962	chelem	0,888	aune	0,752
découpé	0,873	circonstance	0,949	obtient	0,898	viol	0,916	coût	0,880	vêtement	0,869	compagne	0,843	arme	0,757	folie	0,661	alias	0,961	Ittifak	0,886	lune	0,750
revendiquant	0,865	contradictoire	0,948	initient	0,896	poids	0,912	informatique	0,880	abruptement	0,862	discipline	0,843	frappe	0,757	magie	0,654	Jennifer	0,959	Cooke	0,886	dune	0,750
excellent	0,864	journalistique	0,948	dénient	0,893	pain	0,911	outil	0,880	dépècement	0,860	base	0,841	physique	0,755	malades	0,643	Edelmann	0,958	emparé	0,885	lagune	0,739
maître	0,861	motivation	0,947	abstient	0,892	degré	0,910	fiscale	0,877	gratuitement	0,855	réunification	0,840	incapacité	0,753	maladroit	0,635	Jim	0,958	trophée	0,884	l'une	0,733
Prizren	0,861	mollah	0,947	prient	0,892	jazz	0,910	collectif	0,874	tristement	0,855	caisse	0,835	favorise	0,753	malnutrition	0,632	Perez	0,957	triomphe	0,883	d'une	0,723

FIGURE 3 – 10 Mots les plus proches de 4 mots : patient, traitement, maladie et jaune par 3 approches différentes pour le corpus non médical.

D'après la Figure 3, on remarque que les mots les plus proches des 4 mots cibles, peu importe l'approche n'ont que très peu de lien entre eux. On a donc voulu tracer cela dans la Figure 4 graphiquement pour voir ce que cela donne visuellement. (Le script utilisé pour cette visualisation graphique est : Viz_press_wordembeddings.py).

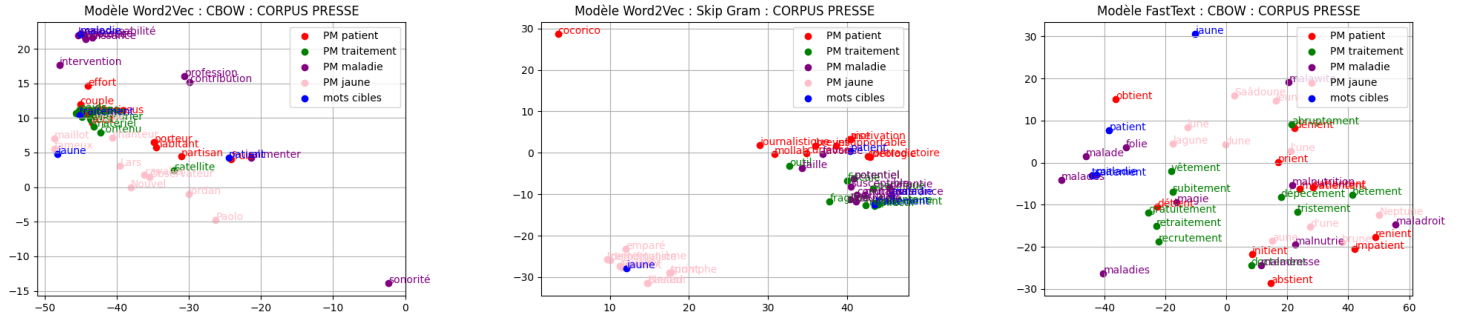

FIGURE 4 – Représentation graphique des mots les plus proches pour 3 approches sur un corpus non médical.

On remarque graphiquement que pour l'approche FastTex, les clusters de mots les plus proches ne se distinguent pas bien. Pour l'approche Word2Vec CBOW on voit une légère répartition en diagonale des clusters. Pour Word2Vec Skipgram par contre la répartition des clusters est plus nette notamment pour le mot "jaune" qui se distingue des 3 autres mots.

Ainsi on remarque que l'approche Word2Vec SkipGram a tendance a donner de meilleure résultats pour notre jeu de donnée que l'approche Word2Vec CBOW et que l'approche FastText.

2.3 Comparaison des embeddings (même approche) entrainés sur deux corpus différents (médical et non médical)

Nous souhaitons maintenant comparer les 3 approches pour les deux jeux de données : médical et non médical (presse)

2.3.1 Word2Vec avec CBOW

Nous récapitulons les résultats obtenus des 10 mots les plus proches pour 4 mots cibles sur les corpus médicaux et non médicaux avec l'approche Word2Vec CBOW sur le Tableau suivant (Figure 5) et le graphique suivant (Figure 6).

w2v CBOW																
patient				traitement				maladie				jaune				
MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		
soigneusement	0,986	malaise	0,888	par	0,968	poumon	0,933	directement	0,986	résolution	0,866	LA	0,997	Michael	0,972	
lorsqu	0,982	consensus	0,883	exemple	0,968	coût	0,925	ADN	0,985	responsabilité	0,864	DU	0,997	Joey	0,965	
allergique	0,979	souffle	0,881	positif	0,958	contenu	0,922	la	0,985	puissance	0,860	Etude	0,997	Spadea	0,964	
signe	0,979	emprunt	0,877	arrêt	0,956	lycée	0,918	rare	0,985	perspective	0,851	EN	0,997	maillot	0,962	
confirmé	0,977	diplôme	0,874	diagnostic	0,956	cancer	0,917	Crohn	0,982	constitution	0,849	DE	0,997	Kahn	0,962	
devra	0,977	découpé	0,873	confirmé	0,954	viol	0,916	embolique	0,981	compagne	0,843	DES	0,996	alias	0,961	
particulièrement	0,974	revendiquant	0,865	comité	0,953	poids	0,912	groupe	0,981	discipline	0,843	syndrome	0,996	Jennifer	0,959	
Il	0,974	excellent	0,864	antérieur	0,951	pain	0,911	prudence	0,980	base	0,841	lamivudine	0,996	Edelmann	0,958	
agir	0,972	maître	0,861	affections	0,950	degré	0,910	fin	0,980	réunification	0,840	Lilly	0,996	Jim	0,958	
aptitude	0,971	Prizren	0,861	TYSABRI	0,947	jazz	0,910	levure	0,980	caisse	0,835	32	0,995	Perez	0,957	

FIGURE 5 – 10 mots les plus proches de 4 mots cibles par une approche Word2VC CBOW

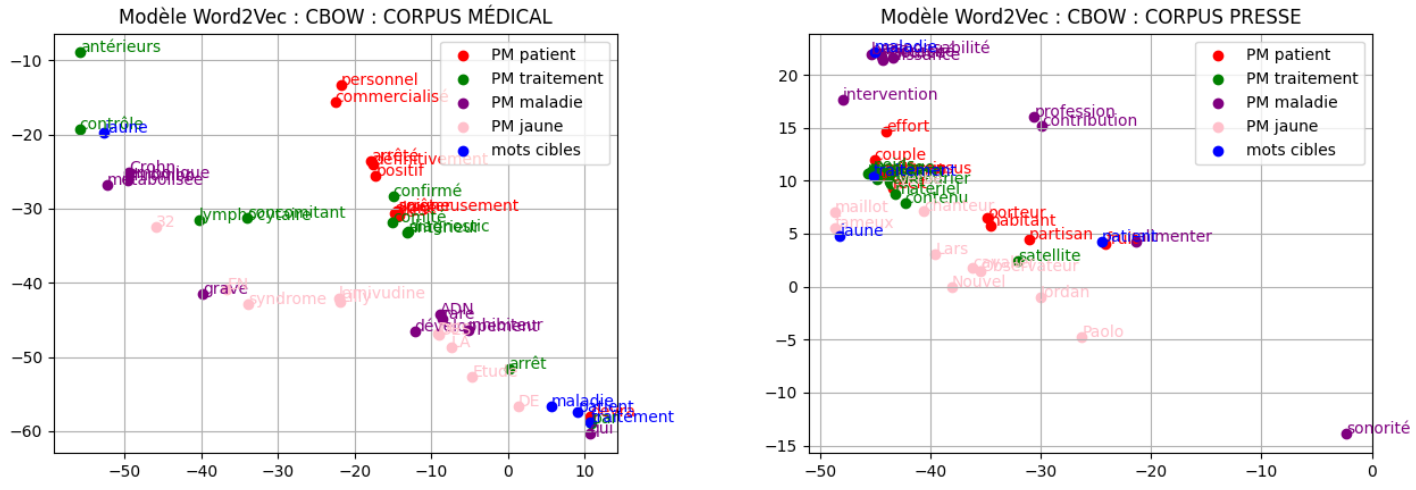

FIGURE 6 – Représentation graphique - Word2Vec CBOW - Medical VS non Médical

2.3.2 Word2Vec avec Skip-Gram

Nous récapitulons les résultats obtenus des 10 mots les plus proches pour 4 mots cibles sur les corpus médicaux et non médicaux avec l'approche Word2Vec Skip-gram (Figure 7) et le graphique suivant (Figure 8).

w2v Skipgram																
patient				traitement				maladie				jaune				
MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		
souffre	0,847	aise	0,956	concomitant	0,790	alimentaire	0,897	thrombo	0,902	assurance	0,870	blanc	0,997	maillot	0,948	
aptitude	0,847	complément	0,953	début	0,788	mépris	0,890	grave	0,902	susceptible	0,782	valeurs	0,996	Bastad	0,912	
modifier	0,842	clown	0,951	Arrêt	0,784	financement	0,887	atteint	0,899	bénéficie	0,761	couleur	0,996	SaintMarin	0,897	
devrait	0,829	insupportable	0,951	mois	0,783	destiné	0,884	thromboembolique	0,891	potentiel	0,761	polyéthylène	0,996	Rafael	0,890	
conscient	0,828	irréaliste	0,949	expérimenté	0,782	potentiel	0,881	Crohn	0,888	garantie	0,759	méthyle	0,996	chelem	0,888	
personnel	0,825	circonstance	0,949	Traitement	0,781	coût	0,880	embolique	0,882	arme	0,757	créatinine	0,996	Ittifak	0,886	
présente	0,824	contradictoire	0,948	antérieurs	0,775	informatique	0,880	considérée	0,882	frappe	0,757	HDPE	0,996	Cooke	0,886	
analgésique	0,822	journalistique	0,948	résultat	0,770	outil	0,880	malgré	0,872	physique	0,755	tumeur	0,996	emparé	0,885	
allergique	0,822	motivation	0,947	définitif	0,769	fiscale	0,877	rare	0,870	incapacité	0,753	phosphate	0,996	trophée	0,884	
remarquer	0,820	mollah	0,947	antérieur	0,769	collectif	0,874	moelle	0,853	favorise	0,753	points	0,996	triomphe	0,883	

FIGURE 7 – 10 mots les plus proches de 4 mots cibles par une approche Word2VC Skip Gram

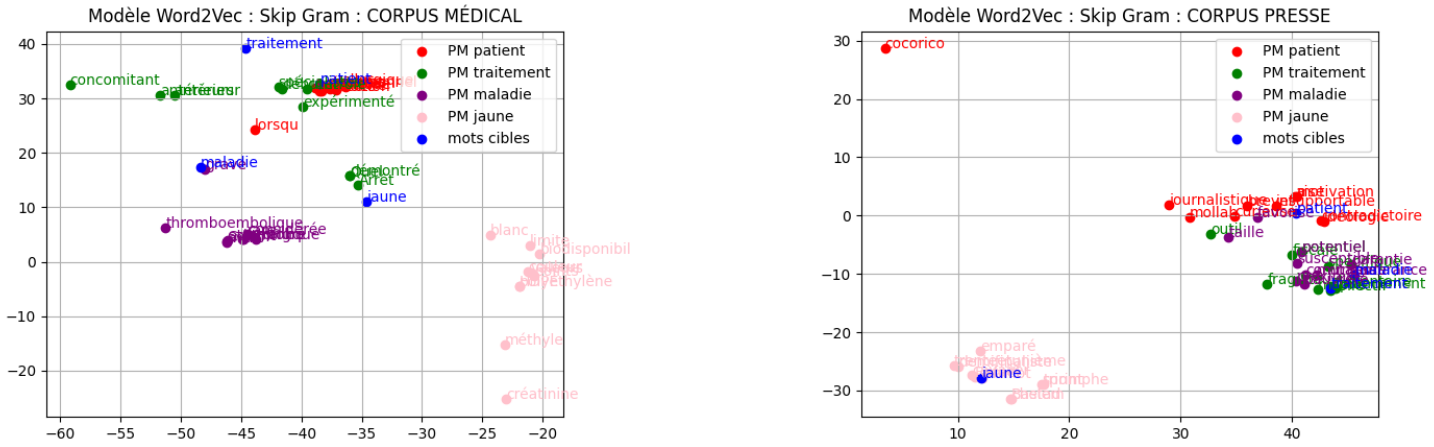

FIGURE 8 – Représentation graphique - Word2Vec CBOW - Medical VS non Médical

2.3.3 FastText avec CBOW

Nous récapitulons les résultats obtenus des 10 mots les plus proches pour 4 mots cibles sur les corpus médicaux et non médicaux avec l'approche FastText CBOW (Figure 9) et le graphique suivant (Figure 10).

FastText CBOW																
patient				traitement				maladie				jaune				
MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		
Patient	0,987	patientent	0,945	Traitement	0,987	retraitement	0,918	malade	0,980	malade	0,787	hexane	0,931	brune	0,798	
patiente	0,961	impatient	0,940	Taaitement	0,978	subitement	0,888	Maladie	0,972	maladies	0,767	goutte	0,923	Jeune	0,783	
parvient	0,948	détient	0,926	traitment	0,973	recrutement	0,884	maldi	0,951	malawite	0,700	gêne	0,921	Neptune	0,778	
appartient	0,945	impatientent	0,909	Allaitement	0,973	bêtement	0,873	malgré	0,919	malnutrie	0,693	titane	0,906	Saâdoune	0,756	
maintient	0,945	renient	0,903	allaitement	0,968	doctement	0,872	malaise	0,918	maladresse	0,672	banane	0,904	aune	0,752	
recevaient	0,928	obtient	0,898	étroitement	0,967	vêtement	0,869	amantadie	0,885	folie	0,661	Rhône	0,902	lune	0,750	
patients	0,918	initient	0,896	évitement	0,965	abruptement	0,862	maligne	0,877	magie	0,654	perchlorure	0,899	dune	0,750	
gradient	0,915	dénient	0,893	entrainement	0,951	dépècement	0,860	malin	0,877	malades	0,643	Hygiène	0,898	lagune	0,739	
excipient	0,915	abstient	0,892	département	0,949	gratuitement	0,855	intolérance	0,858	maladroit	0,635	jeune	0,895	l'une	0,733	
passant	0,915	prient	0,892	recrutement	0,948	tristement	0,855	Parkinson	0,852	malnutrition	0,632	Lane	0,895	d'une	0,723	

FIGURE 9 – 10 mots les plus proches de 4 mots cibles par une approche FastText CBOW

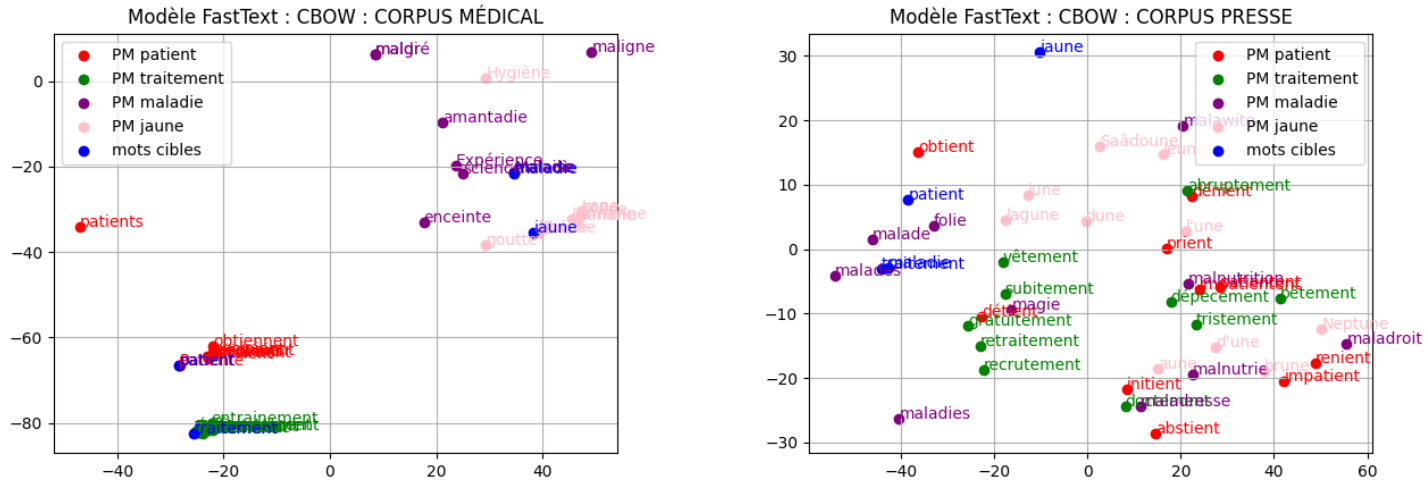

FIGURE 10 – Représentation graphique - Word2Vec CBOW - Medical VS non Médical

D'après les graphiques, on remarque que les clusters sont plus visibles et donc que l'on obtient de meilleurs résultats dans un corpus plus long mais hors du domaine d'étude que dans un corpus plus cours mais dans le domaine d'étude.